An Ensemble Approach for Expanding Queries

Duy Bui BS[2], Doug Redd MS[1,2], Thomas Rindflesch PhD[3], Qing Zeng-Treitler PhD[1,2]

[1]VA Salt Lake City Health Care System; [2]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT; [3]National Library of Medicine, Bethesda, MD

1. Introduction

In our TREC participation, we used an ensemble approach in query expansion. Query expansion, such as synonym expansion, had shown promising results in medical literature search. On the other hand, some of the 2011 papers reported worse results from expansion. Since there are multiple knowledge sources available and each resource has clear strengths and weaknesses, we tested the combination of three expansion methods versus each individual method.

We found that the ensemble approach performed better (in terms of average infAP, infNDCG, R-prec, and P10) than the individual methods and better than the Lucene baseline. The individual expansion methods, however, did not improve the baseline Lucene performance. We also performed an unofficial run using a concept index to boost the query performance, which led to small improvements in infAP, infNDCG, and R-prec.

2. Background

In our own previous studies we have found query expansion to have varying impact depending on the specific queries and the corpus being searched. In one study synonym and predication expansion improved F-measure and recall but reduced precision (1). In a subsequent study they improved MAP scores, but only topic model improved P10. Importantly, precision was not significantly reduced (2). An ensemble method was used in the second study that is different from the one used in this study. That ensemble method used semantic distance to combine the scores of the other methods, but overall it underperformed the individual methods. In this study we used a different ensemble method that uses a simpler summing of weights rather than semantic distance.

3. Methods

In this TREC experiment, our method involved a number of steps with query expansion being the key step (Figure 1).

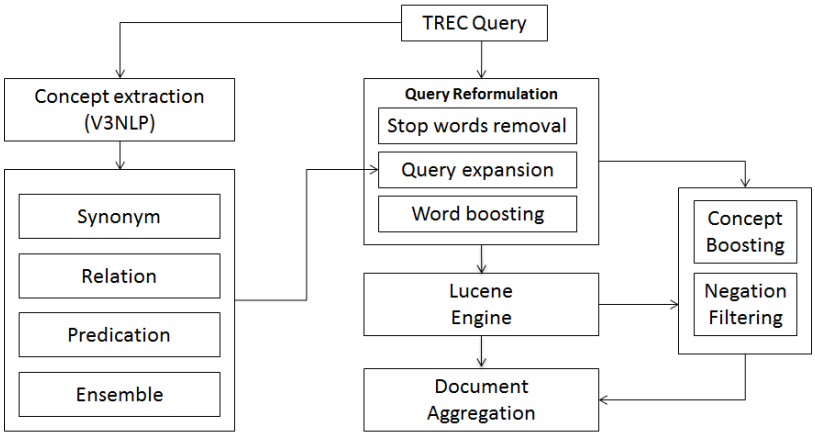

Figure 1 Overview of the query process.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
An Ensemble Approach for Expanding Queries	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Utah,Department of Biomedical Informatics,Salt Lake City,UT,84112**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the Twenty-First Text REtrieval Conference (TREC 2012) held in Gaithersburg, Maryland, November 6-9, 2012. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA). U.S. Government or Federal Rights License

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**8**	

Concept extraction

For each query, we first used V3NLP (3), a concept extraction tool, to identify segments of text that can be mapped to medical concepts.

Query reformulation

Our system then preprocesses and modifies the original query through stop word removal, query expansion, and word level boosting.

Stop word removal: We removed words in our stop word list from the queries. Our stop word list is a merged list of common English stop words and high frequency words in document collection (Table 1).

Table 1. Stop words.

High frequency words		Common English stop words						
but	from	a	can't	he	it's	same	they'll	where
treatment	clear	about	cannot	he'd	its	shan't	they're	where's
normal	patient	above	could	he'll	itself	she	they've	which
him	patients	after	couldn't	he's	let's	she'd	this	while
who	if	again	did	her	me	she'll	those	who
after	R	against	didn't	here	most	she's	through	who's
over	mg	all	do	here's	mustn't	should	to	whom
time	left	am	does	hers	my	shouldn't	too	why
any	do	an	doesn't	herself	myself	so	under	why's
year	all	and	doing	him	no	some	until	with
right	it	any	don't	himself	nor	such	up	won't
had	pressure	are	down	his	not	than	very	would
how	your	aren't	during	how	of	that	was	wouldn't
did	No	as	each	how's	on	that's	wasn't	you
blood	up	at	few	i	once	the	we	you'd
without	history	be	for	i'd	only	their	we'd	you'll
present	which	because	from	i'll	or	theirs	we'll	you're
will	MEDICAL	been	further	i'm	other	them	we're	you've
day	take	before	had	i've	ought	themselves	we've	your
does	an	being	hadn't	if	our	then	were	yours
D	there	below	has	in	ours	there	weren't	yourself
home	also	between	hasn't	into	ourselves	there's	what	yourselves
other	HOURS	both	have	is	out	these	what's	
about	x	but	haven't	isn't	over	they	when	
well	out	by	having	it	own	they'd	when's	

Query expansion: In this study we experimented with three expansion methods plus an ensemble method that incorporated the results of the other three. The three methods were synonym expansion, relation expansion, and predication expansion.

Synonyms:
In this study, we define synonymous terms as the set of concept names belonging to the same unique concept according to a controlled vocabulary. Synonym expansion has been used in previous studies with varying success. Positive results have been obtained using UMLS synonyms when restricting results to the MeSH vocabulary (4). We have implemented a similar approach by using UMLS synonyms with a restricted set of source vocabularies.

We identified synonyms using a combination of tools from the UMLS. We first attempted to map query terms to UMLS concepts using MetaMap, restricting mapping to those scoring -1000 and coming from one of the data sources SNM, NOMEDCT, MSH, or ICD. If this was unsuccessful then we used a term-to-concept lookup table (derived from the MRCONSO table in UMLS, with non-informative terms removed) to find matching concepts. We then used our MRCONSO derived term-to-concept table to identify all terms for the concepts and used those terms as synonyms.

Relations:
In this study, we define related terms as the set of concept names belonging to the same concept where that concept has a relation to the query term's concept according to a controlled vocabulary. Relations used for query expansion have also shown promise (5). Some relations are more informative than others, so in addition to restricting the set of source vocabularies we assigned weights to the different categories of relations.

For related terms, query terms were first mapped to UMLS concepts using MetaMap and the term-to-concept table in the same manner as for synonyms. We then queried related concepts from the metathesaurus MRREL table. We ranked related concepts by summing their weights, with weights assigned by relation category. We assigned weights of: 2 to *child* relationships; 0 to *not related, no mapping*, *allowed qualifier*, and *can be qualified by* relationships; and 1 to all remaining relationship categories.

Predications:
Aside from vocabulary-defined relations, related terms can be identified from other sources such as predication in published medical literature. In this study, we used SemRep which is an NLP system which identifies predications in biomedical documents (6). It builds on the MetaMap application, SPECIALIST lexicon, and Xerox part-of-speech tagger to assign semantic types to noun phrases and identify lexical variants. Evaluation studies have shown its precision to be approximately 75%. We identified predications from a sample corpus where the object was one of the query terms and used the subject of those predications for expansion terms.

To derive expansion terms from predications, we used a database of predications that was generated by running the SemRep program on 10 years of MEDLINE citations (1999-2009). We used the subjects of predications where the object of the predication was a query term. We ranked the subjects by frequency to obtain a ranked list of expansion terms.

Ensemble:
We used an ensemble method to incorporate synonym, relation, and predication expansion into a single result set. Our technique was to normalize the scores from the other methods to values between 0 and 1 using the equation:

$$\text{if } Score(C_x, C_y, s) > 0 : i_s(C_x, C_y, s) = \frac{\ln\left(Score(C_x, C_y, s)\right)+1}{\ln\left(MAX(Score(C_x, C_n, s))\right)+1}$$

$$\text{if } Score(C_x, C_y, s) = 0 : i_s(C_x, C_y, s) = 0$$

where
C_x is the query term
C_y is a related term
C_n is any related term
s is the source of the expansion

The final rank for each expansion term was obtained by summing the scores from the 3 methods. Examples of the expansion results can be seen in Table 2.

Table 2 Examples of synonym, relation, predication, and ensemble expansions

Original query	children	peripheral neuropathy
Synonym expansion	child, child of, offspring, progeny, kid, childhood age person, child youth, offsprings, human child	peripheral nervous system disorders; peripheral nerve diseases; peripheral neuropathies; peripheral nervous system disorder; peripheral nervous system disease; peripheral nerve disease; peripheral nerve disorders, peripheral nerve disorder
Relation expansion	offspring, child of, of child, child find, offsprings, child, progenis, kid, kids, progeny	Neuromuscular disease or syndrome; a-50 myoneural disorders; neuromyopathies; disorders neuromuscular; disease neuromuscular; myoneural disorder, unspecified; neuromuscular disorder; myoneural disorder; myoneural disorders, unspecified; neuromuscular dis
Predication expansion	asthma, adhd, obese, autism, cerebral palsy, disease, symptoms, epilepsy, obesity, overweight	Paclitaxel; bortezomib; vincristine; thalidomide; painful; cisplatin; oxaliplatin; charcot-marie-tooth disease; drugs; neuropathy
Ensemble expansion	child of, asthma, kids, of child, kid, child, adhd, obese, autism	paclitaxel; disease neuromuscular; disorders neuromuscular; myoneural disorders; a-50 myoneural disorders; peripheral nervous system disorders; bortezomib; vincristine; thalidomide; peripheral nerve diseases

We reformulate the query by appending expansion tail right after the concept term. The expansion tail contains a list of recommendation terms connected by the operator OR. We currently set a limit of ten recommendation terms per concept term expansion. An example below demonstrates the expansion tail for the concept term "lupus nephritis" in query number 145.

lupus nephritis (nephritis OR lupus lupus OR glomerulonephritis mycophenolate OR mofetil glomerulonephritis OR lupus cyclophosphamide membranous OR lupus OR nephritis OR syndrome diffuse OR lupus OR glomerulonephritis OR syndrome sle OR membranous OR glomerulonephritis membranous OR lupus OR glomerulonephritis mmf)

Boosting weight for rare words**:** Testing on the 2011 TREC queries, we observed from the previous year (2011) TREC queries containing content words (e.g. medication name) tend to lead to higher precision while those with functional words (in the medical context) such as "developed" or "receiving" lead to many false positives. We also observed that content words tend to have lower frequency, though this is not always true. By assigning each word the inverse document frequency (IDF) as their boosting weight, we gave infrequent words more weight than frequent words.

$$Inverse\ document\ frequency = log \frac{\#Total\ documents}{Document\ frequency}$$

Table 3 Words and their boosting weights.

Original word	Frequency	IDF	Boosted word
Receiving	1891	1.73	receiving^1.73
Procedure	14367	0.85	procedure^0.85
Pain	41292	0.39	pain^0.39
Hospital	15094	0.82	hospital^0.82
Miscarriage	45	3.35	miscarriage^3.35
Radiotherapy	53	3.28	radiotherapy^3.28
Hypoaldosteronism	3	4.53	hypoaldosteronism^4.53
Thyrotoxicosis	20	3.7	thyrotoxicosis^3.7

Lucene Search

We used Lucene version 3.6.1 as the base search engine to test our query expansion methods. We created a Lucene index from the one hundred thousand TREC documents. Query results by Lucene include document IDs and scores in range [0 1] resulting from cosine similarity calculation.

Negation Filtering

We dedicated a separate run to measure the impact of negation terms (e.g No, not) on the TREC corpus. The idea is to conduct two parallel queries: the original query and the negated query. The negated query is the expansion of the original query with negation terms preceding each word. For example, the negated version of "miscarriage^3.35" includes "no miscarriage"^3.35 and "not miscarriage"^3.35. If a document is the result of both original query and negated query, its score is recalculated by subtracting the original score to negated score.

Concept Boosting

Our last experiment (ConceptBoost) used a concept-based indexing method to modify the ranking list of ensemble method. An UMIA-based NLP pipeline Sophia to map medical terms into UMLS CUIs which are then indexed the cuis using Lucene (7, 8). For example, the term "Hepatitis C" is mapped to CUI C0019196. Traditional word based method adds 2 word "Hepatitis" and "C" to the index dictionary while concept method only requires indexing CUI "C0019196" as one entry in concept index. We removed from index the negated concepts detected by Sophia. All queries are translated to CUIs using the sample pipeline, and those CUIs were searched against concept index.

We used concept queries to boost the results of ensemble method. Similar to "Negation Filtering" method, we construct two parallel queries, and the intersected documents of two lists are added score by 1 to push them up to the top of ranked list.

Aggregation

We tried three ways of aggregating (MAX, SUM, AVERAGE) document scores into visit scores. When testing on TREC 2011 data, we observed that using the MAX function yielded higher performance. Hence, we used MAX aggregation in our submission.

4. Evaluation

We submitted 4 automatic runs to the TREC 2012 Medical Records Track. The BMIUOUsyn used a synonym expansion method. The BMIUOUbase is the baseline evaluation using the default Lucene engine using our own stop word list. The BMIUOUens run uses the ensemble expansion methods. Lastly, BMIUOUensneg uses the ensemble method with very simple negation filtering.

The results we received used 4 evaluation metrics: inferred average precision (infAP), inferred normalized discounted cumulated gain (infNDCG), Precision at 10(P@10), and R-precision (R-prec). infAP measures average precision while taking into account the incompleteness of relevance judgments..

Using the relevance judgment set provided by TREC, We extended the evaluation to compare individual expansion method (Predication, Synonym, Relation). We also performed a run without reformulation (LuceneBase) as control.

Table 4 Specification of all experiments

Method	Stop words Removal	Word boosting	Negation Filtering	Concept Boosting	Query Expansion			
					Ensemble	Synonym	Predication	Relation
LuceneBase								
BMIUOUbase*	x							
Synonym	x	x				x		
Predication	x	x					x	
Relation	x	x						x
BMIUOUens*	x	x			x	x	x	x
BMIUOUensneg*	x	x	x		x	x	x	x
ConceptBoost	x	x		x	x	x	x	x

The * indicates the experiment results are submitted to TREC for the official evaluation.

5. Results

Overall, our ensemble and negation methods did better than the median and Lucene baseline (Table 5). Original query with stop word removal (BMIUOUbase) didn't perform as good as the median. We also report original query without stop word removal in the evaluation extension section. Our best submission is the ensemble with negation filtering method although there is little difference from the ensemble without negation.

Table 5 Comparison average results of all official submissions.

	infAP	infNDCG	R-prec	P10
ConceptBoost	0.2001	0.4447	0.3254	0.4660
BMIUOUensneg*	0.1761	0.4275	0.3044	0.4809
BMIUOUens*	0.1754	0.4256	0.3024	0.4787
Median	0.1695	0.4243	0.2935	0.4702
BMIUOUbase*	0.1663	0.4139	0.292	0.4468
Synonym	0.1636	0.3981	0.2783	0.4064
LuceneBase	0.1521	0.3875	0.2785	0.4213
Relation	0.1592	0.3836	0.2697	0.3809
Predication	0.1396	0.3589	0.2598	0.3702

The * indicate the experiment results are submitted to TREC for the official evaluation.

We also counted the number of queries for each method that did better, equal or worse than the median. In table 6, for instance, the ensemble method performed better in 25 queries than the median (infAP), but worse in 19 queries. The individual expansion methods performed worse than our Lucene baseline while the ensemble method performed better than the baseline. We also note that removing stop words improved the Lucene default performance.

Using parallel query for reordering ranked list contributes to overall performance. While the "Negation Filtering" impact is insignificant, the use of concept indexing to boost traditional word based query is showing significant improvement.

Table 6. Count numbers of queries our methods are better, equal or worse than the median of all submissions.

		BMIUOUbase	BMIUOUens	BMIUOUensneg
infAP	Better	21	25	25
	Equal	4	3	3
	Worse	22	19	19
infNDCG	Better	22	23	23
	Equal	3	2	3
	Worse	22	22	21
R-prec	Better	18	24	24
	Equal	15	10	10
	Worse	14	13	13
P@10	Better	12	19	19
	Equal	20	15	15
	Worse	15	13	13

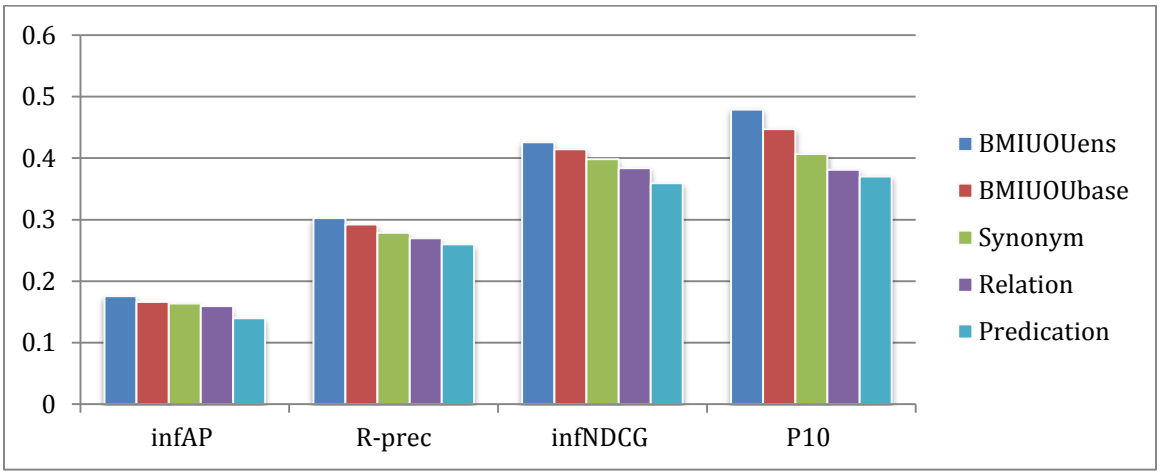

Figure 2. Compare Ensemble method with its three constituent methods.

6. Discussion

Although Lucene is a powerful search engine, our ensemble method did perform better than our baseline Lucene runs. Particularly worth noting is that each individual expansion method did not perform very well. However, combining the methods did show improvements. When we tested our methods on the 2011 TREC, we observed somewhat different results: the ensemble as well as individual expansion methods performed better on bpref but worse on P10 than the baseline Lucene.

Our methods often didn't perform well in queries involving age-specific references. For example, some queries (136, 141,164, 169,170,173,174, and 175) contain keywords: Children, Adults, Elderly. Our expansion term methods are not very helpful when applied to these keywords. Since age is typically available as structured data, we did not develop special functions to handle these keywords.

The concept boosting method has demonstrated beneficial effect on retrieval performance. Concept-based search alone do not outperformed word-based search. This can be attributed to the fact that concept extraction tools are imperfect and do miss some concepts. Despite this, a document retrieved by both concept and word-indexing methods has a higher chance of being the relevant document. The experiment with concept boosting has improved infAP, infNDCG, and R-prec in compare with ensemble method by a couple of percentage points.

7. Acknowledgements

This work was funded by VA grants CHIR HIR 08-374 and VINCI HIR 08-204.

References

1. Zeng QT, Redd D, Rindflesch T, Nebeker J. Synonym, Topic Model and Predicate-Based Query Expansion for Retrieving Clinical Documents. AMIA 2012 Annual Symposium; Nov. 5, 2012; Chicago, IL(In Press).
2. Redd D, Rindflesch TC, Nebeker J, Zeng-Treitler Q. Improve Retrieval Performance on Clinical Notes: A Comparison of Four Methods. Hawaii International Conference on System Sciences; Maui, Hawaii2013.
3. Sasisekharan R, Seshadri V, Weiss SM. Data mining and forecasting in large-scale telecommunication networks. IEEE Expert. 1996;11(1):37-43.
4. Griffon N, Chebil W, Rollin L, Kerdelhue G, Thirion B, Gehanno J-F, et al. Performance evaluation of unified medical language system®'s synonyms expansion to query PubMed. BMC Medical Informatics and Decision Making. 2012;12(12).
5. Zeng QT, Crowell J, Plovnick RM, Kim E, Ngo L, Dibble E. Assisting consumer health information retrieval with query recommendations. J Am Med Inform Assoc. 2006;13(1):80-90. Epub 2005/10/14.
6. Rindflesch TC, Bean CA, Sneiderman CA, editors. Argument identification for arterial branching predications asserted in cardiac catheterization reports. Proc AMIA Annual Symposium; 2000 2000.
7. Divita G, Zeng-Trietler Q, editors. Finding medically unexplained symptoms within VA clinical documents using v3NLP. International Society for Disease Surveillance Conference 2010; 2010.
8. FERRUCCI D, LALLY A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering. 2004;10(3-4):327-48.